

# Video Object Detection and Segmentation Based on Proposal Boxes

Xiaodi Zhang, Zhiguo Cao<sup>(✉)</sup>, Yang Xiao, and Furong Zhao

National Key Laboratory of Science and Technology on Multi-Spectral Information  
Processing, School of Automation,  
Huazhong University of Science and Technology, Wuhan, China  
{xiaodizhang, zgcao, Yang\_Xiao, zhaofr639}@hust.edu.cn

**Abstract.** In this paper, we propose a new method to detect and segment foreground object in video automatically. Given a video sequence, our method begins by generating proposal bounding boxes in each frame, according to both static and motion cues. The boxes are used to detect the primary object in the sequence. We measure each box with its likelihood of containing a foreground object, connect boxes in adjacent frames and calculate the similarity between them. A layered Directed Acyclic Graph is constructed to select object box in each frame. With the help of the object boxes, we model the motion and appearance of the object. Motion cues and appearance cues are combined into an energy minimization framework to obtain the coherent foreground object segmentation in the whole video. Our method reports comparable results with state-of-the-art works on challenging benchmark dataset.

**Keywords:** Video analysis · Video object segmentation · Object proposals

## 1 Introduction

Video object segmentation is one of the fundamental problems in video analysis. It aims at separating the foreground object from its background in a video sequence. This technique is beneficial to a variety of applications, such as video summarization, video retrieval and action recognition.

Humans have the talent to distinguish an object from its background in static images or videos. However, it is difficult for a computer to accomplish such task. Finding or segmenting objects in images is one of the core problems in the field of computer vision. Many exciting approaches have been explored for this type of task, such as figure-ground segmentation, object proposal techniques, saliency detections and object discovery. When we process video data, the task of object discovery and segmentation changes a lot since continuous frames provide motion information. Therefore, both appearance and motion cues should be considered to find out the primary object.

Our approach aims at video object segmentation, while we also give out category independent object detection results in the form of bounding boxes.

Similar to some methods [1–3], we utilize an object proposal method to obtain the initial locations of the object. However, the former methods need to pre-segment each frame and measure the likelihood of each segment to be an object. This process costs a lot of time (more than 1 min per frame). Different from these methods, we generate proposal bounding boxes instead of object-like segments, and it is much more efficient (about 1 s per frame). Both static and motion cues are under consideration for the boxes generation. We attempt to measure the proposal bounding boxes with two scores: objectness score and motion score. Objectness score reports the likelihood of containing an object, and motion score estimates the motion difference between the box and its surrounding area. Since the object always moves differently from its background, the box which has the object inside it oughts to own high objectness score and motion score. In most situations, object moves smoothly across frames in a video, its location and appearance vary slowly. Therefore, the proper object boxes in the consecutive frames are also coherent in location and size. We connect two boxes in consecutive frames and measure the similarity between them. A layered Directed Acyclic Graph (DAG) is constructed to formulate the bounding boxes in the whole sequence, and the problem of selecting boxes is transformed into finding out the path with highest score in the layered DAG. When the boxes are determined, the foreground object is detected. With the help of the selected boxes, we locate the object in each frame, and model the motion and appearance of the object according to the regions inside the boxes. The final segmentation is performed in an energy minimization framework.

The rest of this paper is organized as follows. Section 2 of this paper reviews the related researches on the task of video object segmentation. In Sect. 3, our approach is introduced and discussed in detail. The experimental results and analysis are reported in Sect. 4. The paper is concluded in Sect. 5.

## 2 Related Work

Lots of methods have been explored to fulfill the task of video object segmentation. Divided by the need for manual annotation, the methods are summarized as semi-automatic manner and fully automatic manner. Semi-automatic methods require manual annotations of object position in some key frames for initialization, while the latter scenario doesn't need any human intervention. Without any priori knowledge of the foreground object, the fully automatic methods have to firstly answer the questions of what and where the object is. Different strategies have been developed for the questions. Trajectory analysis and object proposal techniques are often used to discover the object in videos.

**Semi-automatic Methods.** Some semi-automatic methods [4–7] require annotation of precious object segments in key frames. The segments are propagated to other frames under the constrains of motion and appearance. Other semi-automatic methods [8] require object location in the first frame as initialization.

These methods track the object regions in the rest frames. Semi-automatic methods usually get better result than fully automatic methods as they obtain prior knowledge about object annotated by interaction. However, since labor cost is expensive, these methods are unsuitable for large-scale video data processing.

**Trajectory Based Methods.** The main characteristic of trajectory based methods [9–11] is that they analyze long term motion over several frames rather than forward/backward optical flow. These methods assume that trajectories of moving object are similar with each other and different from background. Brox *et al.* [9] defined a distance between trajectories as the maximum difference of their motion over time. Given the distance between trajectories, an affinity matrix was built for the whole sequence and the trajectories were clustered based on the matrix. Lezama *et al.* [11] combined local appearance and motion measurements with long range motion cues in the form of grouped point trajectories. Fragkiadaki *et al.* [10] proposed an embedding discontinuity detector for localizing object boundaries in trajectory spectral embeddings. Trajectory clustering was replaced by discontinuities detection. These methods suffered from some problems such as model selection of clustering and no-rigid/articulated motion.

**Proposal Based Methods.** Object proposals are regarded as regions or windows likely to be an object in an image. Object proposal techniques [1–3] try to find out an object in the image based on bottom-up segmentation. These techniques are beneficial for video object segmentation task as they segment frames in advance. One of the most important steps in these methods is to distinguish which proposals are the regions of the object. Proposal based methods generate hundreds of proposals and measure the probabilities to be a foreground object using appearance and motion cues. After that video object segmentation transforms into a proposal selection problem. Lee *et al.* [12] preformed spectral clustering on proposals to discover the reliable proposals in the key frames. The proposal cluster with highest average score was corresponding to the primary foreground object. These proposals were used to generate object segments in rest frames. Although this method outperformed some semi-automatic methods, the main drawback was that clustering abandoned temporal connections of the proposals and the object-regions were only obtained in some key frames. Zhang *et al.* [13] designed a layered Directed Acyclic Graph to solve proposal selection problem. The layered DAG considers temporal relationship of proposals in consecutive frames and the problem of proposals selection transformed to get the longest weighted path in the DAG. When the path was determined, the most suitable proposal was selected in each frame. Our method is inspired by their brilliant idea. Perazzi *et al.* [14] formulated proposals in a fully connected manner and trained a classifier to measure the proposal regions. Proposal based methods report state-of-art result in this task. They also handle no-rigid and articulated motion well since a prior segmentation is performed in each frame without the influence of motion. However they face the problem of high computational complexity for generating proposal regions. It costs several minutes to

get hundreds of proposal segments in a standard scaled image. The unacceptable time cost limits practical value of these methods.

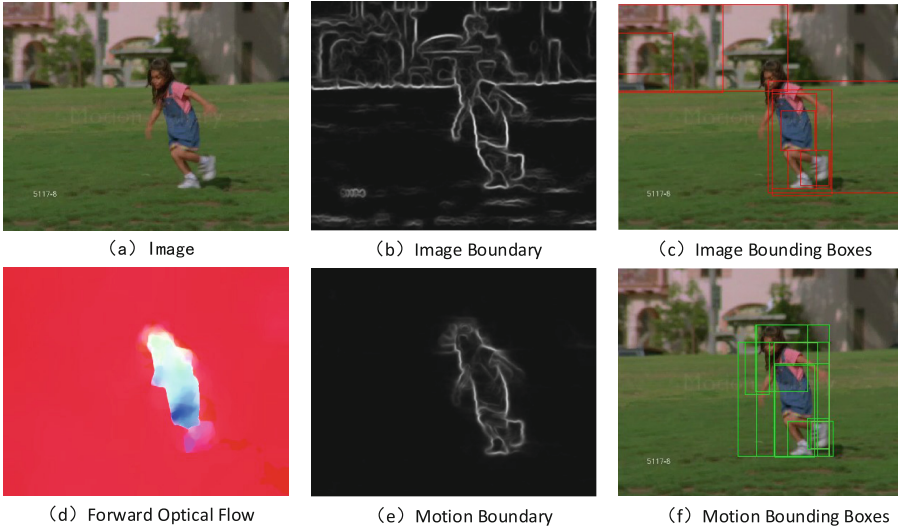
Apart from the aforementioned paradigms, several other formulations have been explored for this task. Papazoglou *et al.* [15] proposed a novel solution, which combined motion boundaries and point-in-polygon problem for efficient initial foreground estimation. The method reported comparable results to proposal based methods while being orders of magnitude faster. Wang *et al.* [16] proposed a saliency-based method for video object segmentation. This method firstly generated framewise spatiotemporal saliency maps using geodesic distance. The object owned high saliency value in the maps and was easy to be segmented.

### 3 Our Approach

Our proposed approach is explained detailedly in this section. There are three main stages in the approach: (1) Proposal bounding boxes generation in each frame; (2) Layered DAG construction and bounding box selection; (3) Modeling motion and appearance for the foreground object and obtaining the final segmentation in an energy minimization framework. The forward optical flow between the  $t$ -th and  $(t + 1)$ -th frame is computed using [17] previously.

#### 3.1 Proposal Bounding Boxes Generation

We utilize the efficient object proposal technique Edge Boxes [18] to generate proposal bounding boxes in each frame. Edge Boxes [18], as its name implies, measures edges in an image, returns hundreds to thousands bounding boxes along with their objectness scores. In order to obtain reliable bounding boxes of the foreground object, we extract two types of edges, one is image edges extracted by structured random forests [19], the other one is motion boundary extracted by [20] which improves the structured random forests [19] and makes it work for motion boundary detection. Both frame image and its forward optical flow are used to extract the motion boundary. Utilizing the edges, Edge Boxes [18] generates image bounding boxes and motion bounding boxes and their objectness scores. Notice that the object may be motionless and the optical flow may be inaccurate in some frames, which leads to false motion boundary and bounding boxes. However, we can always get high-quality image bounding boxes only if there are enough image boundaries. Figure 1 demonstrates the process of bounding boxes generation. As the illustration shows, in this case, the motion bounding boxes in (f) are more concentrated around the girl than the image bounding boxes in (c), due to the clear motion boundary in (e). But when the optical flow is inaccurate, motion bounding boxes turn to be unreliable. Figure 2 demonstrates a failing case of optical flow. The motion boundary fails following the inaccurate optical flow, and then the motion bounding boxes are outside of the target. However, the image bounding boxes are not influenced. We choose 100 bounding boxes with the highest scores from each type, and merge them to be the candidate regions of the object.



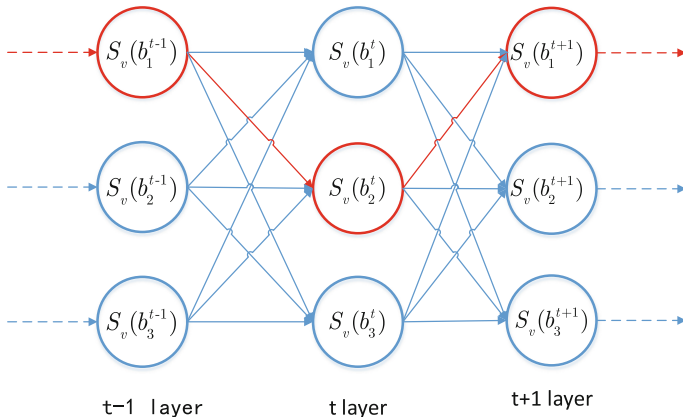
**Fig. 1.** (a) One of the input frames. (b) Image boundary extracted by structured random forests [19]. (c) Some top ranked bounding boxes using image boundary. (d) Forward optical flow of the image. (e) Motion boundary of the image extracted by [20]. The origin image and the forward optical flow are used to extract the motion boundary. (f) Some top ranked motion bounding boxes.



**Fig. 2.** (a) A failing case of optical flow. (b) Motion boundary. The motion boundary fails following the optical flow. (c) Some top ranked motion bounding boxes. The motion bounding boxes are outside of the target. (d) Some top ranked image bounding boxes.

### 3.2 Layered DAG for Bounding Box Selection

We have obtained hundreds of bounding boxes for each frame. It is difficult to determine which bounding box is best for the object in a single frame since the bounding boxes provide candidate regions of the object. Therefore, we consider the consistency of the boxes in the sequence. As the object moves smoothly across the frames, the box associated with it also moves. We want to obtain the boxes which have high probabilities to contain an object tightly and move coherently across the frames. A layered DAG is constructed to formulate the motion of the boxes. The problem of selecting the best box for each frame transforms into



**Fig. 3.** Structure of layered DAG. Circles represent vertices of the graph. The vertices in consecutive frames are connected by straight lines in pairs. The straight lines represent edges of the graph. The path with maximum weight is selected by dynamic programming method. Red circles and lines are parts of the selected path. (Color figure online)

finding the longest path in the layered DAG. Figure 3 demonstrates the structure of the layered DAG. Each frame is represented by a layer in the graph, and each bounding box in the frame turns to be a vertex of the corresponding layer. Boxes in adjacent frames are directly connected in pairs by the edges of the DAG. All the vertices and the edges are weighted, which will be reviewing later on.

**Vertices.** We measure every bounding box with its potential of containing the foreground object. The scores of all bounding boxes are used as the vertices.  $S_v(b_n^t)$  represents score of the  $n$ -th bounding box in the  $t$ -th frame. As both appearance and motion cues are useful for determining moving object in video,  $S_v(b_n^t)$  is made of two parts as:

$$S_v(b_n^t) = A_v(b_n^t) + M_v(b_n^t), \quad (1)$$

where  $A_v(b_n^t)$  is the objectness score of  $b_n^t$  obtained by Edge Boxes [18], and  $M_v(b_n^t)$  is the motion score.  $M_v(b_n^t)$  measures the moving difference between  $b_n^t$  and its surrounding area.  $M_v(b_n^t)$  is defined according to the optical flow histogram:

$$M_v(b_n^t) = 1 - \exp(-\chi_{flow}^2(b_n^t, \bar{b}_n^t)), \quad (2)$$

where  $\chi_{flow}^2(b_n^t, \bar{b}_n^t)$  is the  $\chi^2$ -distance between  $L_1$ -normalized optical flow histograms,  $\bar{b}_n^t$  is a set of pixels within a larger box and around  $b_n^t$ .

**Edges.** Bounding boxes in consecutive frames are connected in pairs to form the edges of the layered DAG. Edges measures the similarity between two connected boxes. As object moves smoothly across frames, object bounding box

also changes coherently in location and size. Therefore, we measure the similarity score of the boxes in two consecutive frames by their locations, sizes and overlap ratio. The similarity score of the boxes is used as edges in the layered DAG and is defined as:

$$S_e(b_t, b_{t+1}) = \lambda * S_g(b_t, b_{t+1}) * S_o(b_t, b_{t+1}), \quad (3)$$

where  $b_t$  and  $b_{t+1}$  are two boxes from  $t$ -th frame and  $(t+1)$ -th frame,  $S_g(b_t, b_{t+1})$  is location and size similarity, and  $S_o(b_t, b_{t+1})$  is overlap similarity.  $\lambda$  is a balance factor between  $S_e$  and  $S_v$ .  $S_g(b_t, b_{t+1})$  and  $S_o(b_t, b_{t+1})$  are defined as:

$$S_g(b_t, b_{t+1}) = \exp\left(-\frac{\|g_t - g_{t+1}\|_2}{h_t + w_t}\right), \quad (4)$$

$$S_o(b_t, b_{t+1}) = \frac{|b_{t+1} \cap \text{warp}(b_t)|}{|b_{t+1} \cup \text{warp}(b_t)|}. \quad (5)$$

In Eq. 4,  $g_t = [x, y, w, h]$  is location and size of box  $b_t$ , where  $[x, y]$  is centroid coordinate of box and  $[w, h]$  is width and height respectively. In Eq. 5,  $\text{warp}(b_t)$  is the warped region from  $b_t$  to frame  $t + 1$  by the optical flow.

**Box Selection.** A proper object bounding box in a certain frame is considered to own high objectness score and motion score, while proper boxes in the consecutive frames are close to each other, or high similarity score  $S_e$  in other words. Once the layered DAG is constructed, the path with maximum total score in the graph represents the most suitable boxes in the frames. This problem can be solved by dynamic programming in linear complexity. The vertices in the maximum weighted path represent object boxes in each frame. After box selection, we obtain a set of object boxes and only one box is left in each frame.

### 3.3 Object Segmentation

We oversegment frames into superpixels by the algorithm SLIC [21] for less computational complexity. After all these, video object segmentation is formulated as a superpixel labeling problem with two labels (foreground/background). Each superpixel in the video sequence takes a label  $l_i^t$  from  $\mathbf{L} = \{0, 1\}$  where 0 represents background and 1 represents foreground. Similar to the related works [12, 13, 15, 16], we define an energy function for the labeling problem:

$$E(l) = \sum_{t,i} A(l_i^t) + \alpha_1 \sum_{t,i} M(l_i^t) + \alpha_2 \sum_{(i,j,t) \in N_s} V(l_i^t, l_j^t) + \alpha_3 \sum_{(i,j,t) \in N_t} W(l_i^t, l_j^{t+1}), \quad (6)$$

where  $A(l_i^t)$  is an appearance unary term and  $M(l_i^t)$  is a motion unary term associated with a superpixel.  $V(l_i^t, l_j^t)$  and  $W(l_i^t, l_j^{t+1})$  are pairwise terms associated with spatial and temporal consistency respectively.  $N_s$  is a set of spatial neighborhoods of a superpixel and  $N_t$  is a set of temporal neighborhoods. A superpixel is warped to the next frame by forward optical flow, and the superpixels

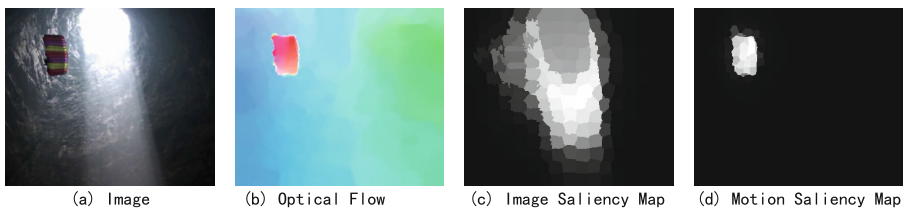
in the next frame which is overlapped with the warped region are considered as its temporal neighborhoods.  $i, j$  are indexes of superpixels. Unary terms try to determine labels of superpixels by appearance and motion cues, while pairwise terms insure spatial and temporal coherence of the segmentation.  $\alpha_1, \alpha_2$  and  $\alpha_3$  are balance coefficients. Equation 6 is optimized via graph-cuts model [22] efficiently.

**Motion Term.** Saliency detection is always used as a technique to discover salient object in images. But saliency detection in video sequence always fails to get a satisfying result, as the primary object in video may not be salient in visualization. However, since object moves differently from its background, it will be salient in optical flow fields. We turn optical flow to RGB image by visualizing it, and use saliency detection method [23] to get motion saliency map. [23] provides a robust and efficient method for visual saliency detection. Figure 4 compares image saliency map with its motion saliency map. The image saliency map Fig. 4(c) fails to get the object (parachute), while it is outstanding in the motion saliency map (d). The motion term in Eq. 6 is defined as:

$$M(l_i^t) = \begin{cases} -\log(1 - S^t(x_i^t)) & l_i^t = 0; \\ -\log(S^t(x_i^t)) & l_i^t = 1. \end{cases} \quad (7)$$

where  $S^t(x_i^t)$  is the motion saliency value of superpixel  $x_i^t$ .

**Appearance Term.** Two Gaussian Mixture Models (GMM) in RGB color space are estimated to model the appearance of foreground and background respectively. We select a part of superpixels and separate them into two sets of fg/bg. We set two conditions for a superpixel to be a member of foreground: (1) inside the selected object box; (2) its motion saliency value is larger than mean value of the frame. Superpixels outside the box with lower motion saliency values are regarded as background superpixels. After that we obtain two sets of superpixels from the whole sequence. Mean RGB colors of the superpixels are used to estimate GMMs for both foreground and background. The appearance term  $A(l_i^t)$  is the negative log-probability of  $x_i^k$  to take label  $l_i^t$  under the associated GMM.



**Fig. 4.** (a) One frame in sequence *parachute*. (b) Forward optical flow of (a). (c) Image saliency map of (a). (d) Motion saliency map of (a) and (b).



**Pairwise Terms.**  $V(l_i^t, l_j^t)$  and  $W(l_i^t, l_j^{t+1})$  are standard contrast-modulated Potts potentials, and follow the definition in [24]:

$$V(l_i^t, l_j^t) = \text{dist}(x_i^t, x_j^t)^{-1} [l_i^t \neq l_j^t] \exp(-\beta_1 \text{col}(x_i^t, x_j^t)^2), \quad (8)$$

$$W(l_i^t, l_j^{t+1}) = \varphi(x_i^t, x_j^{t+1}) [l_i^t \neq l_j^{t+1}] \exp(-\beta_2 \text{col}(x_i^t, x_j^{t+1})^2), \quad (9)$$

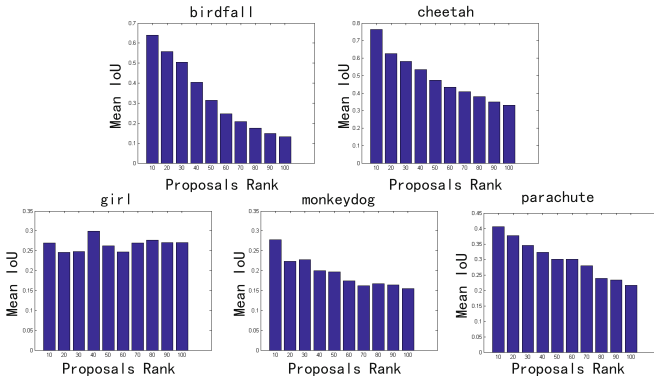
where  $\text{dist}(x_i^t, x_j^t)$  and  $\text{col}(x_i^t, x_j^t)$  are the Euclidean distances between the average positions and average RGB colors of the two superpixels respectively,  $[\bullet]$  is an indicator function,  $\varphi(x_i^t, x_j^{t+1})$  is the overlap ratio of the warped region of  $x_i^t$  and  $x_j^{t+1}$ . The pairwise terms encourage superpixels with close RGB colors to get the same label if they are spatially or temporally connected.

## 4 Experimental Results

We evaluate our approach on SegTrack dataset [7] and FBMS dataset [9]. SegTrack dataset contains 6 videos and pixel-level ground-truth of foreground object in every frame. Following some related works [12, 13, 16], we abandon the sequence *penguin* since there are many penguins moving in the sequence and it is hard to determine which of them is the foreground object. The sequences in this dataset are quite challenging for tracking and video segmentation. *birdfall* has similar colors in fg and bg. There are large camera motion and large shape deformation in *cheetha* and *monkeydog*. *girl* suffers from articulated motion. SegTrack is a benchmark for video object segmentation task. The dataset doesn't supply the object bounding boxes, therefore we manually annotate the box in each frame as ground-truth. We designate the object bounding box as the minimum rectangle that contains the whole object.

There are some parameters in our methods. In Eq. 6, the balance factor  $\lambda$  is fixed as 0.5. For the energy function Eq. 6, we set  $\alpha_1 = 0.4$  and  $\alpha_2 = \alpha_3 = 20$ . In the pairwise terms Eqs. 8 and 9, we set  $\beta_1 = \beta_2 = 1/100$ . The parameters are kept fixed in all experiments.

Firstly, the effectiveness of  $S_v(b_n^t)$  is evaluated. As mentioned previously,  $S_v(b_n^t)$  is designed to measure the probability of a box to be the object bounding box. We rank proposal bounding boxes by descending order of  $S_v(b_n^t)$ , take out the top 100 boxes in each frame and calculate mean IoU of every 10 boxes. Figure 5 reports the ranked proposals-mean IoU results. As the graph shows, in most sequences high rank proposal bounding boxes get high mean IoU. Among the sequences, *girl* reports low and uniform mean IoU value among all the ranks. This is due to the object is large in this sequence, and it is possible to get many boxes that have high overlap ratio with ground-truth. On the other hand, the articulated parts such as legs and hands obtain large motion scores, and corresponding proposal boxes get high rank. Since these boxes lack of consistency, they are not selected as the object boxes. Table 1 reports the mean IoU of the selected boxes in each sequence. *cheetah* gets the lowest mean IoU score because in the beginning frames the box covers two moving objects (cheetah and antelope), since they are close to each other. Although *girl* reports bad results in



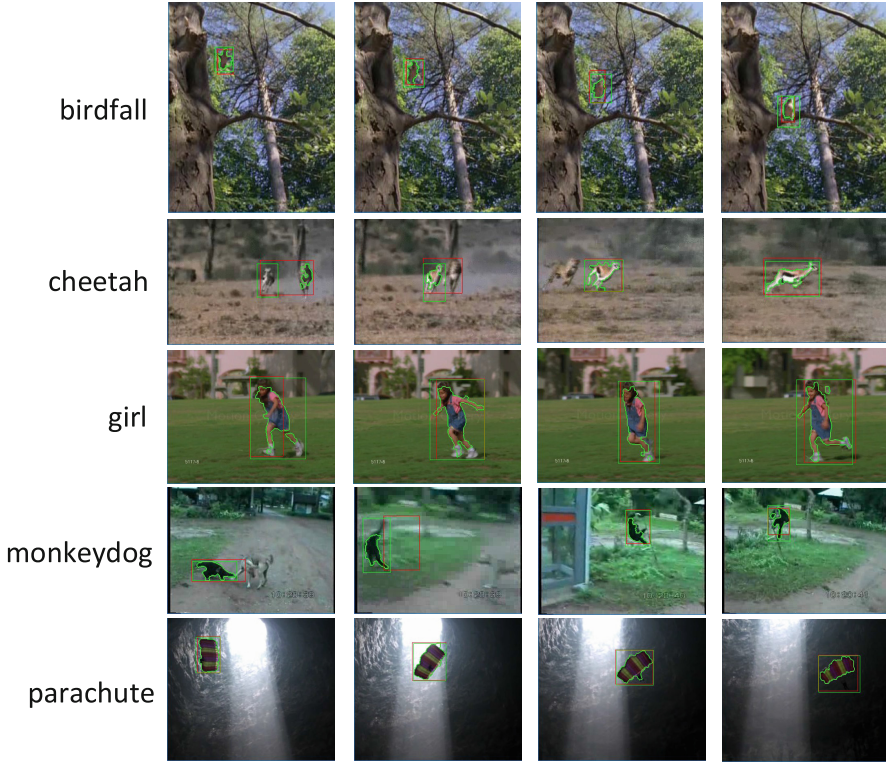
**Fig. 5.** The ranked proposal bounding boxes in different sequences and their mean IoU compared with ground-truth bounding boxes.

Fig. 5, it succeeds to get proper boxes and obtains a high mean IoU in Table 1, due to the consistency of the selected boxes.

**Table 1.** Mean IoU of sequences in SegTrack dataset.

Sequence	Birdfall	Cheetah	Girl	Monkeydog	Parachute
<i>meanIoU</i>	0.688	0.527	0.723	0.627	0.907

Figure 6 demonstrates some results of foreground object detection and the final segmentation on SegTrack dataset [7]. The green rectangles in Fig. 6 are the ground-truth bounding boxes annotated by us, while the red rectangles are the selected bounding boxes. As the illustration shows, most selected boxes catch object tightly and very close to the ground-truth bounding boxes, especially in sequence *parachute* and *birdfall*. In sequence *cheetah*, the antelope is regraded as the primary object in the dataset, while the cheetah also appears and moves in the beginning frames. Our method selects a big bounding box that contains both the antelope and the cheetah at first. When the cheetah disappears, the selected bounding box turns to only contain the antelope. In sequence *monkeydog*, when the monkey is close to boundary of the image, the selected box hasn't followed it. However, when the monkey returns to the middle of the image, the selected box catches it again. The regions in green boundaries in Fig. 6 are the segmented foreground objects. We notice that in *girl*, the foots are usually missed due to the heavy motion blur. In *cheetha*, the cheetah is segmented as the object at the beginning frames since both cheetah and antelope are inside the selected boxes. Table 2 reports quantitative results and comparison with some related works on SegTrack dataset. The results in Table 2 are the average number of mislabelled pixels pre frame compared to the ground-truth. The definition in



**Fig. 6.** Results of object detection and segmentation on SegTrack Dataset. The red rectangles are the selected bounding boxes, and the green rectangles are the manually annotated bounding boxes. The regions within green boundaries are the segmented foreground objects. (Color figure online)

[7] is  $error = \frac{XOR(S,GT)}{F}$ , where  $S$  is the segmentation result,  $GT$  is the ground-truth labeling and  $F$  is the frame number of the sequence. As Table 2 reports, our method gets comparably result with state-of-the-art works.

**Table 2.** Quantitative results and comparison with state-of-the-art works on SegTrack dataset.

Methods	Ours	[13]	[16]	[12]	[15]	[9]	[7]	[25]
Birdfall	183	155	209	288	217	458	252	189
Cheetah	849	633	796	905	890	1968	1142	806
Girl	1943	1488	1040	1785	3859	7595	1304	1698
Monkeydog	501	365	562	521	284	1434	563	472
Parachute	337	220	207	201	855	1113	235	221



**Fig. 7.** Some segmentation results on FBMS dataset. The regions within green boundaries are the segmented foreground objects. (Color figure online)

We also test our approach on FMBS dataset qualitatively. Figure 7 demonstrates some results in this dataset. We notice when the object moves in an articulated manner, our approach may select part instead of the whole object. In *marple1* and *marple3*, pedestrian’s head is detected as the foreground object.

## 5 Conclusion

In this paper, we propose a new approach for the task of video object segmentation. Compared with the former works in this task, our approach works based on proposal bounding boxes and avoids heavy computation for generating segments of proposal regions. The boxes are integrated into a layered DAG, and the problem of box selection can be easily and efficiently solved. With the selected boxes, the foreground object is detected in each frame. The final segmentation is performed based on both motion cues and appearance cues. The experimental results on SegTrack dataset and FBMS dataset testifies the effectiveness of our approach.

**Acknowledgements.** This work was supported by the National High-Tech R&D Program of China (863 Program) under Grant 2015AA015904.

## References

1. Alexe, B., Deselaers, T., Ferrari, V.: What is an object? In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 73–80. IEEE (2010)

2. Endres, I., Hoiem, D.: Category independent object proposals. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010. LNCS, vol. 6315, pp. 575–588. Springer, Heidelberg (2010). doi:[10.1007/978-3-642-15555-0\\_42](https://doi.org/10.1007/978-3-642-15555-0_42)
3. Manen, S., Guillaumin, M., Gool, L.: Prime object proposals with randomized prim’s algorithm. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 2536–2543. IEEE (2013)
4. Bai, X., Wang, J., Simons, D., Sapiro, G.: Video snapcut: robust video object cutout using localized classifiers. *ACM Trans. Grap. (TOG)* **28**, 70 (2009)
5. Price, B.L., Morse, B.S., Cohen, S.: Livecut: learning-based interactive video segmentation by evaluation of multiple propagated cues. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 779–786. IEEE (2009)
6. Yuen, J., Russell, B., Liu, C., Torralba, A.: Labelme video: building a video database with human annotations. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 1451–1458. IEEE (2009)
7. Tsai, D., Flagg, M., Nakazawa, A., Rehg, J.M.: Motion coherent tracking using multi-label MRF optimization. *Int. J. Comput. Vis.* **100**, 190–202 (2012)
8. Ren, X., Malik, J.: Tracking as repeated figure/ground segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1–8. IEEE (2007)
9. Brox, T., Malik, J.: Object segmentation by long term analysis of point trajectories. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010. LNCS, vol. 6315, pp. 282–295. Springer, Heidelberg (2010). doi:[10.1007/978-3-642-15555-0\\_21](https://doi.org/10.1007/978-3-642-15555-0_21)
10. Fragkiadaki, K., Zhang, G., Shi, J.: Video segmentation by tracing discontinuities in a trajectory embedding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1846–1853. IEEE (2012)
11. Lezama, J., Alahari, K., Sivic, J., Laptev, I.: Track to the future: spatio-temporal video segmentation with long-range motion cues. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE (2011)
12. Lee, Y.J., Kim, J., Grauman, K.: Key-segments for video object segmentation. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 1995–2002. IEEE (2011)
13. Zhang, D., Javed, O., Shah, M.: Video object segmentation through spatially accurate and temporally dense extraction of primary object regions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 628–635. IEEE (2013)
14. Perazzi, F., Wang, O., Gross, M., Sorkine-Hornung, A.: Fully connected object proposals for video segmentation. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 3227–3234. IEEE (2015)
15. Papazoglou, A., Ferrari, V.: Fast object segmentation in unconstrained video. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 1777–1784. IEEE (2013)
16. Wang, W., Shen, J., Porikli, F.: Saliency-aware geodesic video object segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3395–3402. IEEE (2015)
17. Sundaram, N., Brox, T., Keutzer, K.: Dense point trajectories by GPU-accelerated large displacement optical flow. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010. LNCS, vol. 6311, pp. 438–451. Springer, Heidelberg (2010). doi:[10.1007/978-3-642-15549-9\\_32](https://doi.org/10.1007/978-3-642-15549-9_32)
18. Zitnick, C.L., Dollár, P.: Edge boxes: locating object proposals from edges. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8693, pp. 391–405. Springer, Heidelberg (2014). doi:[10.1007/978-3-319-10602-1\\_26](https://doi.org/10.1007/978-3-319-10602-1_26)

19. Dollár, P., Zitnick, C.: Structured forests for fast edge detection. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 1841–1848. IEEE (2013)
20. Weinzaepfel, P., Revaud, J., Harchaoui, Z., Schmid, C.: Learning to detect motion boundaries. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2578–2586. IEEE (2015)
21. Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., Susstrunk, S.: SLIC superpixels compared to state-of-the-art superpixel methods. *IEEE Trans. Pattern Anal. Mach. Intell.* **34**, 2274–2282 (2012)
22. Boykov, Y., Veksler, O., Zabih, R.: Fast approximate energy minimization via graph cuts. *IEEE Trans. Pattern Anal. Mach. Intell.* **23**, 1222–1239 (2001)
23. Zhu, W., Liang, S., Wei, Y., Sun, J.: Saliency optimization from robust background detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2814–2821. IEEE (2014)
24. Rother, C., Kolmogorov, V., Blake, A.: Grabcut: interactive foreground extraction using iterated graph cuts. *ACM Trans. Graph. (TOG)* **23**, 309–314 (2004)
25. Ma, T., Latecki, L.J.: Maximum weight cliques with mutex constraints for video object segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 670–677. IEEE (2012)